[Part 1 - Course Intro and Introduction to R](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306" \l "section-9)

This part's material is devided into two subparts.  Part 1A provides an introduction to the course. Part 1B will introduce R. 

#### Part 1A

**Topics:**

**Introduction**(to the course)

* Organisation of the course
* What is Statistics?
* Motivation
* Goals
* Topics and Schedule
* Assessment

#### Part 1B

**Topics:**

**Introduction to R**

* First Steps
  + What is R and why are we going to use it?
  + Where you can get R and some useful accessories
  + Setting it up
  + Help
* Basics of R usage
  + Basic Language Specifics and Commands
  + More Advanced Language Specifics – R Types
  + Common Data Structures
  + Data Anaylsis with R

[Part 2 - Simple Linear Regression](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306#section-10)

**Topics:**

* Descriptive Statistics
* Simple Linear Regression
* Coefficient of Determination

In the first part we started with an introduction to R. R is the statistical software that we are going to use for the data analysis and visualization tasks throughout this course. In this part we will talk about the concepts and techniques that we need in order to carry out a common data analysis that starts with a research objective and works on data that had been collected with respect to that research objective. We are going to cover the concepts in the order of their application in the statistical process.

The tasks that follow the collection of the data are its organisation and a versatile description of it. The concepts and techniques that are attributed to these parts of the statistical process are classified as descriptive statistics. As stated before, participants of the course are expected to have a basic background in descriptive statistics. Still, we will not skip the discussion of descriptive methods completely, but we will cover one particular topic of it: Simple Linear Regression. Since the discussion of this topic will require the knowledge of more basic descriptive concepts, it will serve as a review of the basic methods and indicate where some refreshment is needed. The main reason though for this selection is, that we build a more solid fundament this way for topics that will be covered in a later part of the course.

[Part 3 - Probability and Combinatorics](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306#section-11)

**Topics:**

* Probability Theory (Probability Rules)
* Counting Techniques/Combinatorics
* Probability Distributions  
  − Discrete Probability Distributions

1. Binomial Prob. Distribution
2. Hypergeometric Prob. Distribution

− Continiuous Probability Distributions

* Uniform Prob. Distribution
* Normal Prob. Distribution

The analysis of sample data requires the application of techniques that allow you to take the results from the sample and extend it to the population. The collection of those techniques is refered to as inferential statistics. Probability forms the foundation of inferential statistics. With this part you will get to know the basics of probability theory and combinatorics in order to get prepared for the discussion of the inferential concepts.

[Part 4 - Discrete Probability Distributions](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306#section-12)

**Topics:**

* Random Variables
* Probability Distributions
* Discrete Probability Distributions
* Binomial Distribution
* Hypergeometric Distribution

This week will be about random variables and discrete probability distributions.

[Part 5 - Continuous Probability Distributions](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306#section-13)

**Topics:**

* Continuous Random Variables
* Probability Density Function
* The Uniform Distribution
* The Normal Distribution
* Standard Normal Curve
* Assessing Normality – Normal Probability Plot

[Part 6 - Sampling Distributions](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306#section-14)

**Topics:**

* Sampling Distributions
* Sampling Dist. of the Sample Mean
* Sampling Dist. of the Sample Proportion
* Central Limit Theorem

[Part 7 - Estimation](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306#section-15)

**Topics:**

* Point Estimate, Estimator
* Confidence Interval
* Level of Confidence, Critical Value
* Z-Interval
* t-Interval
* Determining Sample Size

[Part 8 - Statistical Tests](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306#section-16)

**Topics:**

* Hypothesis Test, Null Hypothesis, Alternative Hypothesis, One-Tailed, Two-Tailed
* Type-I Error, Type-II Error
* Level of Significance
* Test Statistic
* One-Sample Z-Test
* One-Sample t-Test
* Critical Value
* P-Value
* Parametric tests vs. Non-Parametric tests
* Statistical vs. Practical Significance

[Part 9](https://moodle.hochschule-rhein-waal.de/course/view.php?id=13306#section-17)

**Topics:**

* Parametric tests vs. Non-Parametric tests
* Dealing with Data Issues, Robust Methods
* Linear Regression

**Data Analysis Process:**

1) Research objective.

2) Data Collection

3) Describe the data.

* clean, organize, summarize, model
* analyze to understand or predict

4) Perform inference.

* generalize sample data

Data Collection:

* what do we measure?
* Which type of study?
* How and when do we measure?
* Which entities do we measure?

Exam Question and solution:

Part:1

1)  Explain the meaning of the following terms in DataAnalysis:

Variable:

An attribute/characteristic that describes a person or object is called variable. The value of the variable can vary from one entity to another or at different times.

Data:

Data are a set of values representing one or more entities (persons or objects->place, thing, or idea). Data can comprise values of one or more variables (->univariate data, resp. multivariate data).

Population:

The entire group of individuals to be studied is called population. An individual is a person or object that is a member of the population being studied. The population of concern must be carefully defined.

Sample:

A sample is a subset of a population.

Parameter  
A parameter is a numerical summary of a population.

Statistic  
A statistic is a numerical summary of a sample.

(2)  In statistics one can distinguish between descriptive statistics and inferential statistics. **Explain what these two fields of statistics are about.**

**Descriptive statistics** is what most people think of, when they hear the word statistics.

The basic subset of statistical methods that allow to organize, summarize, transform, model and present data in an informative way.

Descriptive statistics describe data through models, numerical summaries, tables and graphs.

**Inferential statistics** uses methods that take results from a sample, extend it to the population, and measure the reliability of the result.

Inferential statistics is related to another field of mathematics, probability. The theory of statistics is based on the laws of probability

(3)  A classical problem driven data analysis can be considered a process that can be divided into 4 major steps. Which are those steps.

(4)  ConsiderthefollowingRcode:

1. # Create a vector with integer values
2. valVec <- c(12, 16, 22, 7, 32)
3. # Operate on the vector with the integer values.
4. valVecConv <- valVec \*2/1+1
5. # Print the results to the console.
6. valVecConv
7. Which value or values will be printed to the console (format does not matter)?
8. 25 33 45 15 65

5)  In R NA is a reserved token. What does it stand for? What is used to represent?

NA is a reserved token that stands for Not Available and is used to represent missing values in a set of values.

(6)  Why is the default missing value, NA ,a logical vector? What’s special about logical vectors?(Hint: think about c(FALSE, NA\_character\_).)

NA is not a value though and operations on NA result in NA. The motivation for this rule is simply that if the specification of an operation is incomplete, the result cannot be known and hence is not available.

NA can be used in any vector type.

In case we have missing values in a vector (NAs) and operations on NA result in NA, for any operation we want to perform on the values we need to extract them first. The example on the slide illustrates this problem:

x has six elements, the third and the fifth are NA.

(x <- c(10, 20, NA, 4, NA, 2))

[1] 10 20 NA 4 NA 2

sum(x) / length(x)

[1] NA

(7)  Name the four different systems/groups of objects in R.

Basically we can distinguish four different systems/groups of objects in R

1. Base Types
2. S3 Objects
3. S4 Objects
4. RC (Reference Classes) Objects

(8)  What happens to a factor in R when you modify its levels? Consider the following code:

1. f1 < - factor(letters)
2. levels(f1) < - rev(levels(f1)

"z" "y" "x" "w" "v" "u" "t" "s" "r" "q" "p" "o" "n" "m" "l" "k" "j" "i" "h" "g" "f" "e" "d" "c" "b" "a"

(9) Assign the names of the R data structures listed below the following table to the correct empty cells of the table according to the dimensionality of the data structures and their ability to represent a set of different data types (heterogeneous) or not (homogenous).

|  |  |  |
| --- | --- | --- |
| Dimension | Homogeneous | Heterogeneous |
| 1d | Atomic Vector | List |
| 2d | Matrix | Data Frame |
| nd | Arrary |  |

R has no 0-dimensional, or scalar types. Individual numbers or strings are represented by vectors of length one.

Atomic Vector, Matrix, Array, List, Data frame

(10) Which R data structure can one consider the equivalent of a data matrix of a multivariate data set, with variables of different levels of measurement?

Lists <- Elements can have different types

(11) Consider the following R code snippet:

vecA <- c(1,2,3)

vecB <- c(3,4,5)

data <- data.frame(vecA,vecB)

data[2,]

Which values are printed to the console after the last line of code?

> data

vecA vecB

1 1 3

2 2 4

3 3 5

> data[2,]

vecA vecB

2 2 4

Part2

(12) What is Regression Analysis? What is the difference between a simple and a multiple regression?

Regression is an approach for modelling the relationship between a quantitative response and one or more explanatory variables.

The case of one explanatory variable is called simple regression.

For more than one explanatory variable, the process is called multiple regression.

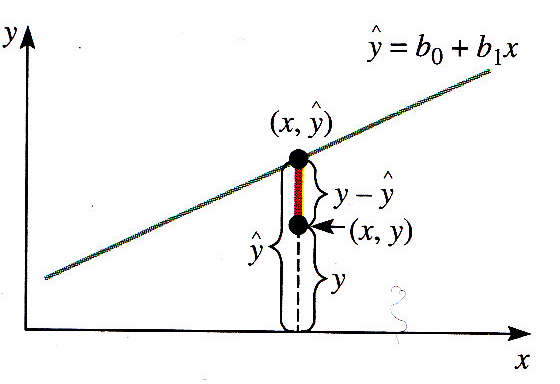
(13) How does the method of the least squares in regression analysis work?

**Solution: Linear Regression: page-25**

If a straight-line model seems appropriate, the best fitting straight line is found by using the method of the least squares, following the least squares

criterion:  
For a line described by   
𝒚 = 𝜷 + 𝜷 𝒙

Find the constants 𝜷 0 and 𝜷 1, such that  
𝒚� = 𝜷 + 𝜷 𝒙



is as small as possible. 𝒚

(14) What does the measure that is called the coefficient of determination (R2) measure?

The coefficient of determination, R2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

PRE – Measure, Proportional Reduction in Error

(15) Consider the following R code (line starts with “>”) and the respective output on the R console:

Continued on next page.

> data<-read.csv("compensation.csv")

> str(data)

'data.frame': 40 obs. of 3 variables:  
$ Root : num 6.22 6.49 4.92 5.13 5.42 ...  
$ Fruit : num 59.8 61 14.7 19.3 34.2 ...  
$ Grazing: Factor w/ 2 levels "Grazed","Ungrazed": 2 2 2 2 2 2 2 2 2 2 ...

> summary(data)

Root Fruit Grazing

Min. : 4.426 Min. : 14.73 Grazed :20

1st Qu.: 6.083

Median : 7.123

Mean : 7.181

3rd Qu.: 8.510

Max. :10.253 Max. :116.05

1st Qu.: 41.15 Ungrazed:20

Median : 60.88

Mean : 59.41

3rd Qu.: 76.19

> model <- lm(data[data$Grazing=="Ungrazed", ]$Fruit~data[data$Grazing=="Ungrazed",

]$Root)

> summary(model)

Call:

lm(formula = data[data$Grazing == "Ungrazed", ]$Fruit ~ data[data$Grazing ==

"Ungrazed", ]$Root)

Residuals:

Min 1Q Median 3Q Max

-9.4542 -4.2430 -0.4643 4.8578 8.8639

Coefficients:

(Intercept)

data[data$Grazing == "Ungrazed", ]$Root

Estimate Std. Error t value Pr(>|t|)

-94.367 9.211 -10.24 6.14e-09 \*\*\*

23.996 1.507 15.93 4.72e-12 \*\*\*

page3image58787712

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.755 on 18 degrees of freedom Multiple R-squared: 0.9337, Adjusted R-squared: 0.93 F-statistic: 253.6 on 1 and 18 DF, p-value: 4.715e-12

* Which particular analysis has been carried out with the lm(...) call?

Answer:

lm is used to fit linear models. It can be used to carry out regression.

* Which variables are modeled?

Answer: Root, Fruit, and Grazing

* What does the resulting model look like?
* Interpret the result of the lm(...)call!
* Visualize(graph) the resulting model?

Part3

(16) For a parallel structure of identical components, the system can succeed if at least one of the components succeeds. Assume that components fail independently of each other and that each component has a 0.15 probability of failure.

-  Would it be unusual to observe one component fail? Two components?

-  What is the probability that a parallel structure with 2 identical components will

succeed?

-  How many components would be needed in the structure so that the probability the system will succeed is greater than 0.9999?

Solution of Problem 16 (Independent Events)

(a) P( one failure)=0.15, not unusual; P(two failure)=0.15×.15=0.0225, unusual (b) P(succeed)= 1-P(fail)=1-0.0225=0.9775

(c) 5 components [as for 5 components probability of failure is 0.15×0.15×0.15×0.15×0.15=.0000759375, So P(succeed)=1-.0000759375=0.999924062]

(17)A box contains 50 red balls and 6 blue balls. What is the probability in a random selection of 5 balls that exactly 3 of the selected ones are blue?

Condition: Randomly selection 5 balls and exactly 3 blue balls

|  |  |  |
| --- | --- | --- |
| Red balls (50) | Blue balls (6) |  |
| 2 | 3 | 50c2X6c3=24500 |

(18) 1% of all people have cancer. 90% of people who have cancer test positive when given a cancer- detecting blood test, meaning the test detects cancer 90% of the time. 5% of people will have false positives, meaning that 5% of the time this test will produce a positive result when people do not have cancer.

Given the above data, what is the probability that a person has cancer if they have a positive cancer-test result? (Note: answers are rounded to the nearest 4th decimal place)

Answer is my notebook.

(19)Data from car accidents all over the country have shown that most accidents occur within 25 miles from the place of residence of the involved drivers. Based on this statistic people might erroneously draw the conclusion that drivers are less likely to have an accident when they are further away from home. Use the given example to explain what the “Confusion of the Inverse” phenomenon is about.

**Part 4**

(20) A term life insurance policy will pay a beneficiary a certain sum of money upon the death of the policyholder.

These policies have premiums that must be paid annually. Suppose a life insurance company offers such a term life insurance policy for 18-year-old males that would pay €150,000 to the beneficiary and would cost €850 annually.

Consider that the probability for an 18-year-old male to survive the year is 0.9998 and compute the expected value of this policy to the insurance company. Interpret the result.

**Answer: Discrete Random Variable**

Two possible outcomes: survival or death  
Random Variable, X: payout depending on the possible outcomes

|  |  |
| --- | --- |
| x | P(x) |
| €850 | 0.9998 |
| -€149150 | 0.0002 |

E(x)= 𝝁𝒙 =ΣxP(x)=850\*0.9998+(-149150\*).0002)= €820

The company expects to make €820 for each 18-year-old male client for insuracne if many policies will be sold.

(21) A shipment of 120 fasteners that contains 4 defective items was sent to a manufacturing plant. The quality control manager at the plant randomly selects 5 fasteners. What is the probability that exactly one of the selected ones is defective?

Now we are going for a different approach:

Classical approach P(E) = N(E)/N(S) with E: 1 of 5 selected fasteners is defective.

We will use counting techniques to determine N(E) and N(S).

All items are considered distinct!

N(S) = Number of possible different arrangements of 120 items, order does not matter: 120C5 = 190,578,024

Number of ways of selecting 1 defective fastener out 4, order does not matter: 4C1

Number of ways of selecting 4 non-defective fasteners out 116, order does not matter: 116C4

N(E) = Number of ways of selecting exactly 1 defective in 5 fasteners:

= 4C1 X 116C4 = 28,640,980

Calculate P(E) = N(E)/N(S): 28,640,980/190,578,024 ≈ 0.15

(22) According to a study 86% of the countries households own a cellularphone (one or more). In a simple random sample of 300 households, 275 owned at least one cellphone. Is this result unusual?

**Answer: Binomial Distribution**

n= 300, p = 0.86

μx = np=258

σx= √np(1-p)=6.0

np(1-p)>9

=300\*0.86(1-0.86)>9

=36.12>9

If not stated otherwise, unusual means a probability less than 0.05. Does that translate here in to the simple question: P (X = 275) < 0.05???

That would make the answer as easy as determining P(X=275) with the PMF (P(x)= nCx px (1-p)n-x ).

P(X=275) = 300C275\*(0.86)^275\*(1-0.86)^300-275

=8.53x10^-04

=0.000853

This is less than 0.05. so, this is unusual.

Part 5

(23) Imagine a friend is usually late. Suppose that she could be equally likely on time or up to 30 minutes late. Given a random appointment with that friend, what is the probability for the friend to be between 10 and 20 minutes late?

**Answer:** Probability Density Function (Page 226, slide: week 4)

(24) The Empirical Rule states that about 68% of the data in a bell-shaped distribution lies within 1 standard deviation of the mean. For the standard normal distribution, this means about 68% of the data lies between z =-1 and z=1. Verify this result. Verify that about 95% of the data lies within 2 standard deviations of the mean. Finally, verify that about 99.7% of the data lies within 3 standard deviations of the mean.

Solution:

P(Z<1)-P(Z<-1)=0.8413-0.1587=0.6826  
So about 68% of the data lies within 1 standard deviation of the mean.

P(Z<2)-P(Z<-2)=0.9772-0.0228=0.9544  
So about 95% of the data lies within 2 standard deviation of the mean.

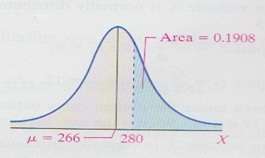
P(Z<3)-P(Z<-3)=0.9987-0.0013=0.9974  
So about 99.7% of the data lies within 3 standard deviation of the mean.

(25) The lengths of human pregnancies are normally distributed with μ= 266 days and σ=16 days. The figure on the left represents the normal curve with μ= 266 days and σ=16 days. The area to the right of x=280 is 0.1908. Provide two interpretations of this area.

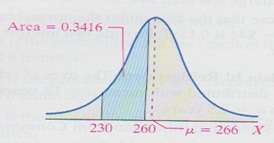
Solution:

(a)(i) The proportion of human pregnancies that last more than 280 days is 0.1908.

(ii) The probability that a randomly selected human pregnancy lasts more than 280 days is 0.1908.



(b)  The figure below represents the normal curve with μ= 266 days and σ=16 days. The area between x=230 and x=260 is 0.3416. Provide two interpretations of this area.



Solution:

(b) (i) The proportion of human pregnancies that last between 230 and 260 days is 0.3416.

(ii) The probability that a randomly selected human pregnancy lasts between 230 and 260 days is 0.3416.

(26) The data in the following table represent the finishing times (in seconds) for six randomly selected races of a greyhound named Barbies Bomber. Is there evidence to support the belief that the variable "finishing time" is normally distributed?

|  |  |
| --- | --- |
| **Barbies Bombers's finishing times** | |
| 31.35 | 32.52 |
| 32.06 | 31.26 |
| 31.91 | 32.37 |

How to get the Normal Score?

1. Arrange the data in ascending order.
2. Compute 𝒇i , the expected proportion of observations less than or equal to the i-th data value:

**fi = (i-0,375)/(n+0,25)**

with i = position of the data value in the ordered list and n= number of observations

3. Find the z-score corresponding to 𝑓𝑓 .

|  |  |  |  |
| --- | --- | --- | --- |
| **Index, i** | **Observed Values** | **fi = (i-0,375)/(n+0,25)** | **Expected z-score** |
| 1 | **31,26** | 0,1 | **-1,28** |
| 2 | **31,35** | 0,26 | **-0,64** |
| 3 | **31,91** | 0,42 | **-0,20** |
| 4 | **32,06** | 0,58 | **0,20** |
| 5 | **32,37** | 0,74 | **0,64** |
| 6 | **32,52** | 0,9 | **1,28** |

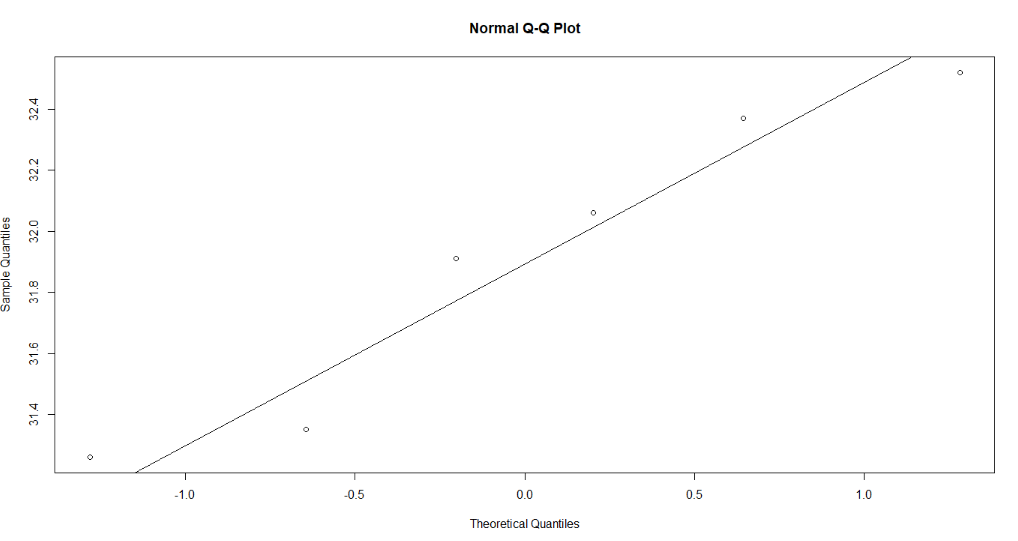
(27) The following lines of code are used in R in order to solve problem (26):

barbVec <- c(31.35,32.52,31.26,32.06,31.91,32.37)

qqBarb <- qqnorm(barbVec)

qqline(barbVec)

As a result R plots the following normal q-q plot:



In brief, explain what this plot shows and which conclusion can be drawn with regard to the problem described in task (26).

Solution:

The normal probability plot in this case does not show major departures from a straight line. A regression line has been added.

The result indicates that the finishing times of the mentioned dog are normally distributed.

Part 6

(28) The length of human pregnancies is approximately normally distributed with mean μ= 266 days and standard deviation σ = 16 days.

- What is the probability a randomly selected pregnancy lasts less than 260 days?

page5image126012608

* -  Suppose a random sample of 20 pregnancies is obtained. Describe the sampling distribution of the sample mean length of human pregnancies.
* -  What is the probability that a random sample of 20 pregnancies has a mean gestation period of 260 days or less?
* -  What is the probability that a random sample of 50 pregnancies has a mean gestation period of 260 days or less?
* -  What might you conclude if a random sample of 50 pregnancies resulted in a mean gestation period of 260 days or less?

Solution: Sampling Distribution, Solution in Week-7, exercise solution

(29) The mean weight gain during pregnancy is 30 pounds, with a standard deviation of 12.9 pounds. Weight gain during pregnancy is skewed right. In a certain neighborhood the mean of a sample with 35 persons is 36.2. Is this result unusual?

Solution: Slide sampling distribution: page-39

(30) 15% of all Americans have hearing trouble.

a) In a random sample of 120 Americans, what is the probability at most 12 % have hearing troubles?

b) Suppose that a random sample of 120 Americans who regularly listen to music using headphones results in 26 having hearing trouble. What might you conclude?

Solution: Slide sampling distribution: page-53

a)

330 million Americans  n is less than 5% of the population

np(1 – p) = 15.3 ≥10

The sampling distribution of the sample proportion is approximately normal with mean 0.15

and standard deviation (approx.) 0.03259 ( 𝑝 ~N(0.15, 0.03259).

P(at most 12% with hearing troubles) = P( 𝑝 ≤ 0.12) ≈ 0.1788

About 18 out of 100 random samples of size 120 will have at most 12% having hearing trouble.

b)

𝑝 = 26/120 = 0.217

P( 𝑝 >= 0.217) ≈ 0.0197

The sample is unusual for a population with 15% having hearing trouble.

Either the proportion value for the population is not right or the sample represents an unlikely selection of persons of the population.

(31) According to the American Red Cross, 7% of people in the United States have blood type 0- negative. If you would want to determine the probability of a single random sample of 500 people in the US with fewer than 30 to have blood type 0-negative, you could get the solution with the following two different code snippets in R.

Code 1:

n <- 500

p <- 0.07

pbinom(29,n,p)

Code 2 (with n\*p\*(1-p)>9):

n <- 500

p <- 0.07

pnorm(29.5, n\*p, sqrt(n\*p\*(1-p)))

Explain the difference between the two and why both give at least approx. the same result.

**Solution: R Script Week-5**

#------------

# Example 8b

#------------

#----------------------------------------------------------

# According to the American Red Cross, 7% of people in the

# United States have blood type 0-negative.

# What is the probability that in a single random sample of

# 500 people in the US fewer than 30 have blood type

# 0-negative?

#----------------------------------------------------------

n <- 500

p <- 0.07

pbinom(29,n,p)

# The answer can also be approximated using the normal dist.

# if the given binomial dist. approx. the shape of

# a normal dist.

# Can we assume that the given binimial distribution is normal

n\*p\*(1-p)>9

# ==> TRUE ==> approx. normal distribution

# Determine the necessary parameters for the normal dist.

# --> Expected Value or Mean of the given binomial dist.

expV\_bloodType <- n\*p

# --> Standard Deviation ?? of the given binomial dist.

sd\_bloodType <- sqrt(n\*p\*(1-p))

# Probability that fewer than 30 have that blood type in a sample

# of 500

pnorm(29.5, expV\_bloodType, sd\_bloodType)

#------------------------------

# Visualisation of the result

#------------------------------

# hist does not work here, since we do not have counts but probabilities/relative frequencies

# we can create a barplot instead, that looks like a histogram

bp <- barplot(dbinom(15:55,n,p), space = 0, width = 1,

#names.arg = as.character(15:55),

col="yellow", main="Binom. Dist. (n=500, p=0.07) with a superimposed normal curve",

axes=FALSE, xlab="No of people with blood typ 0-negative", ylab="Prob. Density")

axis(1, labels = c(15:55), at = bp, tcl = -0.25)

axis(2)

bp

dbinom(15:55,n,p)

# color the interesting area differently

barplot( dbinom(15:29, n, p),

space=0,

col="skyblue",

add=TRUE)

# add a vertical line for x = 29

abline(v = 14.5, col = "red", lty = 1)

text(x=12.5, y=0.06,

pos=4, paste("29"),cex=0.75, col = "red")

xy <- xy.coords(x=0.5:39.5, y= dnorm(15:54,expV\_bloodType,sd\_bloodType))

lines(xy,

lwd=2, col=2)